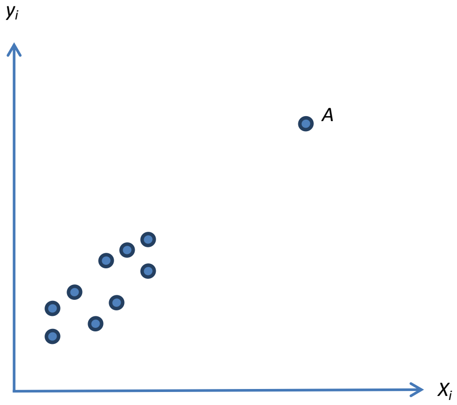
Chapter 6

Diagnostic for Leverage and Influence

The location of observations in x-space can play an important role in determining the regression coefficients. Consider a situation like in the following

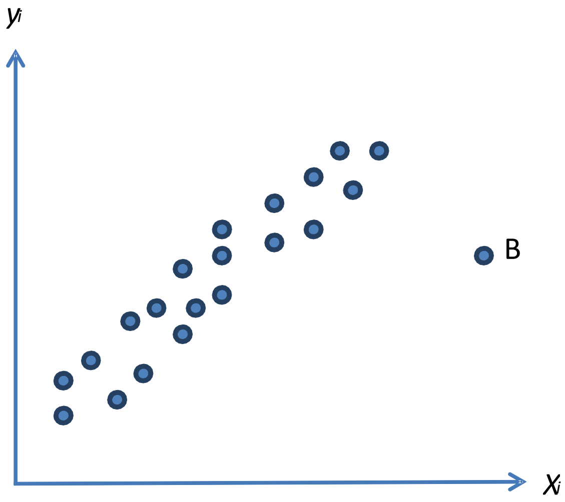


The point A in this figure is remote in $x-$space from the rest of the sample but it lies almost on the regression line passing through the rest of the sample points. This is a **leverage point.**

It is an unusual x-value and may control certain model properties.

- This point does not affect the estimates of the regression coefficients.
- It affects the model summary statistics e.g., R^2, standard errors of regression coefficients etc.

Now consider the point B following figure:



This point has a moderately unusual x-coordinate and the y-value is also unusual. This is an **influential point**

- It has a noticeable impact on the model coefficients.
- It pulls the regression model in its direction.

Sometimes a small subset of data exerts a disproportionate influence on the model coefficients and properties. In an extreme case, the parameter estimates may depend more on the influential subset of points than on the majority of the data. This is an undesirable situation. A regression model has to be a representative of all the sample observations and not only of a few. So we would like to find these influential points and asses their impact on the model.

- If these influential points are “bad” values, they should be eliminated from the sample.
- If nothing is wrong with these points, but if they control the model properties, then it is to be found that how do they affect the regression model in use.

Leverage

The location of points in x-space affects the model properties like parameter estimates, standard errors, predicted values, summary statistics etc. The hat matrix $H = X(X'X)^{-1}X'$ plays an important role in identifying influential observations. Since

$$V(\hat{y}) = \sigma^2 H$$
$$V(e) = \sigma^2 (I - H),$$

($\hat{y}$ is fitted value and e is residual) the elements h_{ii} of H may be interpreted as the amount of leverage excreted by the i^{th} observation y_i on the i^{th} fitted value $\hat{y}_i$.

The i^{th} diagonal element of H is

$$h_{ii} = x_i'(X'X)^{-1}x_i$$

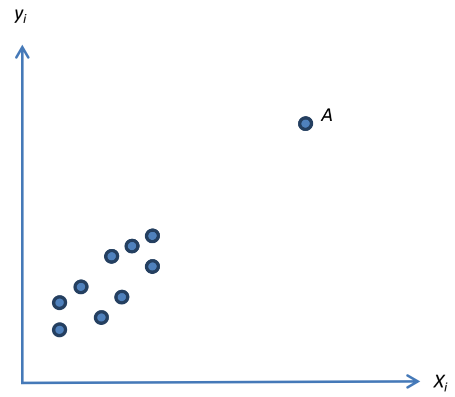
where x_i' is the i^{th} row of X-matrix. The hat matrix diagonal is a standardized measure of the distance of i^{th} an observation from the centre (or centroid) of the $x-$space. Thus large hat diagonals reveal observations that are potentially influential because they are remote in x-space from the rest of the sample.

Average size of hat diagonal $(\bar{h})$ is

$$\begin{aligned}\bar{h} &= \frac{\sum h_{ii}}{n} = \frac{\text{rank}(H)}{n} \\ &= \frac{\text{rank}(X)}{n} \\ &= \frac{trH}{n} \\ &= \frac{k}{n}\end{aligned}$$

- If $h_{ii} > 2\bar{h} = \frac{2k}{n} \Rightarrow$ the point is remote enough from rest of the data to be considered as a leverage point.
- Care is needed in using cutoff value $\frac{2k}{n}$ and magnitudes of k and n are to be assessed. There can be situations where $\frac{2k}{n} > 1$ and then this cut off does not apply.

All leverage points are not influential on the regression coefficients. In the following figure



the point A

- will have a large hat diagonal and is surely a leverage point.
- have no effect of the regression coefficients as it lies on the same line passing through the remaining observations.

Hat diagonal examine only the location of observations in x-space, so we can look at the studentized residual or R-student in conjunction with the h_{ii}.

Observation with

- large hat diagonal and
- large residuals

are likely to be influential.

Measures of influence

(1) Cook's D-statistics:

If data set is small, then the deletion of values greatly affects the fit and statistical conclusions.

In measuring influence, it is desirable to consider both

- the location of point is x-space and
- the response variable.

The Cook's distance statistics denoted as, Cook's D-statistic is a measure of the distance between the least-squares estimate based on all n observations in b and the estimate obtained by deleting the i^{th} point, say $b_{(i)}$. It is given by

$$D_i(M,C) = \frac{(b_{(i)} - b)' M (b_{(i)} - b)}{C}; \quad i = 1, 2, ..., n.$$

The usual choice of M and C are

$$M = X'X$$
$$C = kMS_{res}$$

So

$$D_i(X'X, kMS_{res}) = \frac{(b_{(i)} - b)' X'X (b_{(i)} - b)}{kMS_{res}}; \quad i = 1, 2, ..., n$$
$$= \frac{(\hat{y} - \hat{y}_{(i)})'(\hat{y} - \hat{y}_{(i)})}{kMS_{res}}$$

where

$$\hat{y} = Xb$$
$$\hat{y}_{(i)} = Xb_{(i)}$$
$$b = (X'X)^{-1} X'y.$$

Points with large $D_i \Rightarrow$ the points have considerable influence of OLSE b.

Since

$$\frac{(b_{(i)} - b)' X'X (b_{(i)} - b)/k}{SS_{res}/(n-k)}$$

looks like a statistic having a $F(k, n-k)$ distribution. Note that this statistics is not having a $F(k, n-k)$ distribution.

So the magnitude of D_i is assessed by comparing it with $F_\alpha(k, n-k)$. If $D_i = F_{0.5}(k, n-k),$ then deleting point i would move $\hat{\beta}_{(i)}$ to the boundary of an approximate 50% confidence region for β based on the complete data set.

This displacement is large and indicates that the OLSE is sensitive to the i^{th} data point.

- Since $F_{0.5}(k, n-k) \approx 1,$ we usually consider that points for which $D_i > 1$ to be influential.
- Ideally, each $b_{(i)}$ is expected to stay within the boundary of a 10-20% confidence region.
- D_i is not an F-statistic but cut off of 1 work very well in practice.

Since

$$D_i = \frac{(\hat{y}_{(i)} - \hat{y})'(\hat{y}_{(i)} - \hat{y})}{kMS_{res}},$$

so D_i can be interpreted as a squared Euclidian distance (apart from kMS_{res}) that the vector of fitted values moves when the i^{th} observation is deleted.

Since

$$b_i - b_{(i)} = \frac{(X'X)^{-1}x_i e_i}{1-h_{ii}},$$

the D_i can be written as

$$\begin{aligned} D_i &= \frac{(b_i - b_{(i)})'X'X(b_i - b_{(i)})}{kMS_{res}} \\ &= \frac{x_i'(X'X)^{-1}(X'X)(X'X)^{-1}x_i e_i^2}{(1-h_{ii})^2 kMS_{res}} \\ &= \left(\frac{e_i}{1-h_{ii}}\right)^2 \left(\frac{h_{ii}}{kMS_{res}}\right) \\ &= \frac{r_i^2}{k}\left(\frac{h_{ii}}{1-h_{ii}}\right) \end{aligned}$$

where r_i is studentized residual.

D_i : product of squared i^{th} studentized residual and $h_{ii}/(1-h_{ii})$.

: Reflects how well the model fits the i^{th} observation y_i and a component that measures how far that point is from the rest of the data.

: Either component or both may contribute to a large value of D_i.

: Thus D_i combines residual magnitude for i^{th} observation and location of points in x- space to assess influence.

(2) *DFFITS* and *DFBETAS*:

Cook's distance measure is a deletion diagnostic, i.e., it measures the influence of i^{th} observation if it is removed from the sample.

There are two more statistics:

(i) *DFBETAS* which indicates how much the regression coefficient changes if the i^{th} observation were deleted. Such change is measured in terms of standard deviation units. This statistic is

$$DFBETAS_{j,i} = \frac{b_j - b_{j(i)}}{\sqrt{S_{(i)}^2 C_{jj}}}$$

where C_{jj} is the j^{th} diagonal element of $(X'X)^{-1}$ and $b_{j(i)}$ regression coefficient computed without the use of i^{th} observation.

Large (in magnitude) value of $DFBETAS_{j,i}$, indicates that i^{th} observation has considerable influence on the j^{th} regression coefficient.

(i) The values of $DFBETAS_{j,i}$ can be expressed in a $n \times k$ matrix that conveys similar information to the composite influence information in Cook's distance measure.

(ii) The n elements in the j^{th} row of R produce the leverage that the n observations in the sample have on $\hat{\beta}_j$. $DFBETAS_{j,i}$ is the j^{th} element of $(b - b_{(i)})$ divided by a standardization factor

$$b_i - b_{(i)} = \frac{(X'X)^{-1} x_i' e_i}{1 - h_{ii}}.$$

The j^{th} element of $(b_i - b_{(i)})$ can be expressed as

$$b_i^j - b_{(i)}^j = \frac{r_{j,i} e_i}{1 - h_{ii}}.$$

Let $r_{j,i} = ((R))$ denotes the elements of R, so

$$\begin{aligned}(RR')' &= \left[(X'X)^{-1} X'X (X'X)^{-1}\right]' \\ &= (X'X)^{-1} \\ &= C \\ &= R'R.\end{aligned}$$

Since

$$C_{jj} = r_j^{'} r_j$$

So

$$\sqrt{S_{(i)}^2 C_{jj}} = \sqrt{S_{(i)}^2 r_j^{'} r_j}$$

$$DFBETAS_{j,i} = \frac{b_i^j - b_{(i)}^j}{\sqrt{S_{(i)}^2 C_{jj}}}$$

$$= \left(\frac{r_{j,i} e_i}{1 - h_{ii}} \right) \frac{1}{\sqrt{S_{(i)}^2 r_j^{'} r_j}}$$

$$= \underset{\substack{\downarrow \\ \boxed{\text{Measures leverage (impact of } i^{th} \text{ observation on } b_i)}}}{\frac{r_{j,i}}{\sqrt{r_j^{'} r_j}}} \underset{\substack{\downarrow \\ \boxed{\text{Measures effect of large residuals}}}}{\frac{t_i}{\sqrt{1 - h_{ii}}}}$$

where t_i is the i^{th} R-student residual. Now if $\left| DFBETAS_{j,i} \right| > \frac{2}{\sqrt{n}}$, then i^{th} observation warrants examination.

2. *DFFITS:*

The deletion influence of i^{th} observation on the predicted or fitted value can be investigated by using diagnostic by Belsley, Kuh and Welsch as

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{S_{(i)}^2 h_{ii}}}, \; i = 1, 2, ..., n$$

where $\hat{y}_{(i)}$ is the fitted value of y_i obtained without the use of the i^{th} observation. The denominator is just a standardization, since $Var(\hat{y}_i) = \sigma^2 h_{ii}$.

This $DFFITS_i$ is the number of standard deviations that the fitted value $\hat{y}_i$ changes of i^{th} observation is removed.

Computationally,

$$\begin{aligned} DFFITS_{\text{i}} &= \sqrt{\frac{h_{ii}}{1-h_{ii}}}\frac{e_i}{S_{(i)}\sqrt{1-h_{ii}}} \\ &= t_i\sqrt{\frac{h_{ii}}{1-h_{ii}}} \\ &= R-\text{student} \times \text{leverage of } i^{th} \text{ observation} \end{aligned}$$

where t_i is R-student.

- If the data point is an outlier, then R-student will be large is magnitude.
- If the data point has high leverage, then h_{ii} will be close to unity.
- In either of these cases, $DFFITS_i$ can be large.
- If $h_{ii} \approx 0,$ then the effect of R-student will be moderated.
- If R-student is near to zero, then combined with high leverage point, then $DFFITS_i$ can be a small value.
- Thus $DFFITS_i$ is affected by both leverage and prediction error. Belsley, Kuh and Welsch suggest that any observation for which

  $$\left|DFFITS_{\text{i}}\right| > 2\sqrt{\frac{k}{n}}$$

  warrants attention.

Note: The cutoff values of $DFFITS_{j,i}$ and $DFFITS_i$ are only guidelines. It is very difficult to provide cutoffs that are correct for all cases. So analyst is recommended to utilize information about both what is diagnostic means and the application environment in selecting a cutoff.

For example, if $DFFITS_i = 1$, say, we could translate this into actual response units to determine just how much $\hat{y}_i$ is affected by removing the i^{th} observation.

Then use $DFFITS_{j,i}$ to see whether this observation is responsible for the significance (or perhaps nonsignificance) of particular coefficients or for changes is sign in a regression coefficient.

$DFFITS_{j,i}$ can be used to determine how much change in actual problem-specific units a data point has on the regression coefficient. Sometimes these changes will be of importance in a problem-specific context even though the diagnostic statistic do not exceed the formal cutoff.

The recommended cutoffs are a function of sample size n. Indeed, any formal cutoff should be a function of sample size. However, in practice, these cutoffs often identify more data points than an analyst may wish to analyze. This is particularly true in small samples. The cutoff values provided by Belsley, Kuh and Welsch make more sense for large samples. When n is small, then diagnostic views are preferred.

A measure of model performance generalized variance:

The diagnostics D_i, $DFFITS_{j,i}$ and $DFFITS_i$ provide insight about the effect of observations on the estimated coefficient $\hat{\beta}_j$ and fitted values $\hat{y}_i$. They do not provide any information about the overall precision of estimation.

The **generalized variance** is defined as the determinant of the covariance matrix and is a convenient scalar measure of precision. The generalized variance of OLSE b is

$$GV(b) = \left|V(b)\right| = \left|\sigma^2 (X'X)^{-1}\right|.$$

To express the role of i^{th} observation on the precision of estimation, define

$$COVRATIO_i = \frac{\left|(X'_{(i)}X_{(i)})^{-1}S^2_{(i)}\right|}{\left|(X'X)^{-1}MS_{res}\right|}, \; i = 1, 2, ..., n.$$

If $COVRATIO_i > 1 \Rightarrow i^{th}$ observation improves the precision of estimation.

If $COVRATIO_i < 1 \Rightarrow$ the inclusion of i^{th} observation degrades the precision computationally,

$$COVRATIO_i \;\; = \frac{\left(S^2_{(i)}\right)^k}{MS^k_{res}}\left(\frac{1}{1-h_{ii}}\right) \quad \text{where} \quad \frac{1}{1-h_{ii}} = \frac{\left|(X'_{(i)}X_{(i)})^{-1}\right|}{\left|(X'X)^{-1}\right|}.$$

- So a high leverage point will make $COVRATIO_i$ large. This is logical since a high-leverage point will improve the precision unless the point is an outlier in y-space.
- If i^{th} observation is an outlier, then $\frac{\left(S^2_{(i)}\right)^k}{MS^k_{res}}$ will be much less than unity.
- Cutoff values for $COVRATIO$ are not easy to obtain. It is suggested that

  if $\quad COVRATIO_i > 1 + \frac{3k}{n}$

  or if $\quad COVRATIO_i < 1 - \frac{3k}{n}$,

  then i^{th} point should be considered influential. The lower bound is only appropriate when $n > 3k$.

These cutoffs are only recommended for large samples.